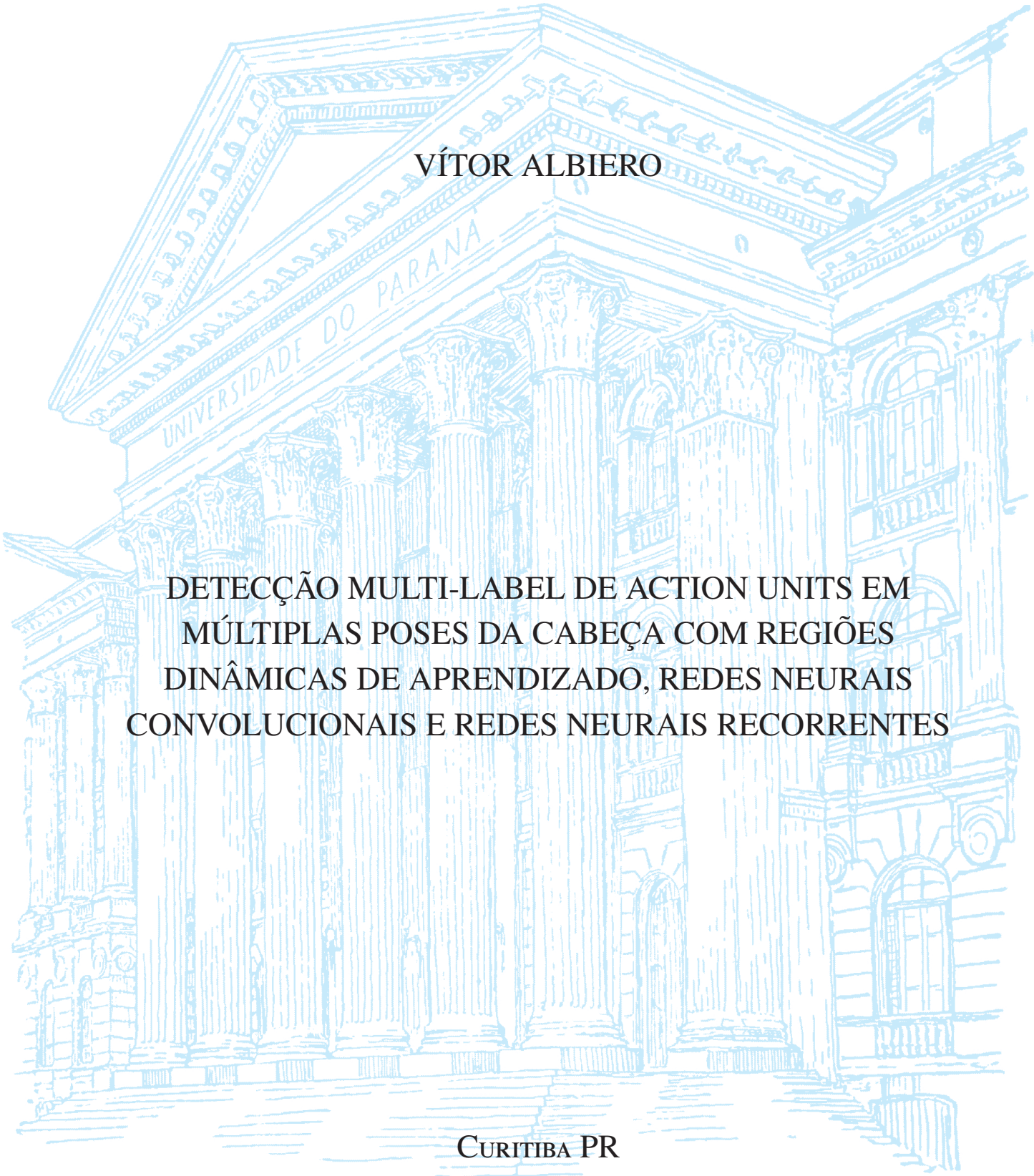
UNIVERSIDADE FEDERAL DO PARANÁ

VÍTOR ALBIERO



DETECÇÃO MULTI-LABEL DE ACTION UNITS EM MÚLTIPLAS POSES DA CABEÇA COM REGIÕES DINÂMICAS DE APRENDIZADO, REDES NEURAIS CONVOLUCIONAIS E REDES NEURAIS RECORRENTES

CURITIBA PR
2018

VÍTOR ALBIERO

DETECÇÃO MULTI-LABEL DE ACTION UNITS EM MÚLTIPLAS POSES DA CABEÇA COM REGIÕES DINÂMICAS DE APRENDIZADO, REDES NEURAIS CONVOLUCIONAIS E REDES NEURAIS RECORRENTES

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Informática no Programa de Pós-Graduação em Informática, setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Prof. Dra. Olga R. P. Bellon.

CURITIBA PR
2018

FICHA CATALOGRÁFICA ELABORADA PELO SISTEMA DE BIBLIOTECAS/UFPR
BIBLIOTECA DE CIÊNCIA E TECNOLOGIA

AL335d Albiero, Vítor
Detecção multi-label de action units em múltiplas poses da cabeça com regiões dinâmicas de aprendizado, redes neurais convolucionais e redes neurais recorrentes / Vítor Albiero. – Curitiba, 2018.
41 p. : il. color.

Dissertação - Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-Graduação em Informática, 2018.

Orientadora: Olga R. P Bellon.

1. Redes neurais. 2. Expressões faciais. 3. Múltiplas poses. I. Universidade Federal do Paraná. II. Bellon, Olga R. P. III. Título.

CDD: 006.32

Bibliotecária: Romilda Santos - CRB-9/1214

MINISTÉRIO DA EDUCAÇÃO
SETOR SETOR DE CIENCIAS EXATAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO INFORMÁTICA

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **VÍTOR ALBIERO** intitulada: **Detecção Multi-label de Action Units em Múltiplas Poses da Cabeça com Regiões Dinâmicas de Aprendizado, Redes Neurais Convolucionais e Redes Neurais Recorrentes**, após terem inquirido o aluno e realizado a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 11 de Julho de 2018.

OLGA REGINA PEREIRA BELLON
Presidente da Banca Examinadora (UFPR)

TEODIANO FREIRE BASTOS FILHO
Avaliador Externo (UFES)

LUCIANO SILVA
Avaliador Interno (UFPR)

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
Setor de Ciências Exatas
Coordenação de Pós-Graduação em Informática

Dedico este trabalho a minha querida esposa Mariany Albiero, que me apoiou de forma incondicional durante todos os momentos do mestrado.

Agradecimentos

À minha querida esposa Mariany Albiero, por todo o auxílio e apoio prestados durante o desenvolvimento deste trabalho. À minha mãe Poliana Albiero e irmão Gustavo Carneiro, por sempre acreditarem em mim.

Aos meus tios Roberto e Sandra Albiero, por me proporcionarem moradia durante o período deste trabalho. Ao meu primo Maurício Albiero, pela convivência, companheirismo e apoio prestado durante o mestrado.

A professora Olga Bellon, orientadora e amiga, que sempre disposta, compartilhou seus conhecimentos e experiências, possibilitando a conclusão deste trabalho. Ao Luciano Silva, professor e amigo, por todo o conhecimento passado durante o mestrado.

Aos colegas do grupo IMAGO, pelo companheirismo e auxílio prestado, em especial: Júlio Cézar Batista, Luan Porfírio, Nathaly Gasparin, Flávio Zavan e Fernando Silva.

Enfim, à todos que de uma forma ou outra colaboraram para que este trabalho fosse realizado com êxito.

Resumo

Este trabalho apresenta a análise de expressões faciais através da detecção *multi-label* de *Action Units* (AUs) em múltiplas poses da cabeça. A análise de expressões faciais em múltiplas poses da cabeça é um problema que detectores robustos de AUs devem lidar, pois é incomum uma pessoa manter sempre a mesma pose ao realizar expressões faciais. Para isto, este trabalho propõe uma abordagem de *region learning* que cria regiões dinâmicas dentro de uma rede neural convolucional (CNN) usando pontos fiduciais faciais. As regiões dinâmicas de aprendizado (DRL) garantem que cada AU esteja no centro da região, assim como siga o movimento da pose da cabeça. A *dynamic region learning* foi implementada no final da rede neural convolucional VGG-Face, utilizado *transfer-learning* para iniciar o treinamento. Além disso, para melhorar as detecções, este trabalho explora informações temporais através de uma rede neural recorrente. Para tal, foi treinada uma rede *Long-Short Term Memory* (LSTM) utilizando características previamente extraídas pela DRL. Os experimentos foram conduzidos na base de dados Facial Expression Recognition and Analysis Challenge (FERA 2017), que contém nove poses diferentes, e mostram que o trabalho proposto foi capaz de se adaptar às noves poses, superando o estado da arte.

Palavras-chave: aprendizado profundo, detecção de unidades de ação, análise de expressões faciais, múltiplas poses, regiões dinâmicas de aprendizado, redes neurais convolucionais, redes neurais recorrentes.

Abstract

This work presents a facial expression analysis through multi-label detection of Action Units (AUs) on multiple head poses. The facial expression analysis on multiple head poses is an issue that robust AU detectors must deal with, as it is uncommon for a person to keep the same pose while performing facial expressions. To this end, this work proposes a region learning approach that creates dynamic regions of interest inside a convolutional neural network (CNN) using facial landmark points. The dynamic region learning (DRL) ensures that each AU is in the center of the region, and also follows the head pose movement. The DRL was implemented in the final part of the VGG-Face convolutional neural network, using transfer-learning to start the training. Also, to improve the detection, this work explores temporal information through a recurrent neural network. For this, a Long-Short Term Memory (LSTM) network was trained using features previously extracted by the DRL. The experiments were conducted on the Facial Expression Recognition (FERA 2017) database, which contains nine different head poses, and shows that the proposed approach was able to adapt to all the head poses, surpassing the state-of-the-art.

Keywords: Deep learning, action units detection, facial expression analysis, multiple head poses, dynamic regions learning, convolutional neural networks, recurrent neural networks.

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

AU	*Action Unit*
CNN	Rede Neural Convolucional
DRL	Região Dinâmica de Aprendizado
FACS	Sistema de Codificação de Unidades de Ação
HOG	*Histogram of Oriented Objects*
LBP	*Local Binary Patterns*
LRCN	*Long-Term Recurrent Convolutional Network*
LSTM	*Long-Short Term Memory*
RNN	Rede Neural Recorrente
UFPR	Universidade Federal do Paraná

Sumário

1 Introdução

O objetivo principal da análise de expressões faciais é, através da visão computacional em conjunto com aprendizado de máquina, determinar a expressão facial de um indivíduo em uma imagem ou vídeo. Inicialmente as expressões faciais foram categorizadas em seis expressões básicas, mais a neutra (quando não há expressão alguma), sendo elas: felicidade, tristeza, raiva, medo, desgosto e surpresa [Ekman et al., 2002]. Porém, conforme [Du et al., 2014], as expressões faciais não são limitadas apenas às básicas, podendo haver combinações entre elas, chamadas de expressões faciais compostas. Até o momento são conhecidas 22 expressões faciais compostas.

Foi proposto por [Ekman et al., 2002] o *Facial Action Coding System* (FACS) que, diferente dos modelos anteriores, não é categorizado em expressões faciais, mas sim em 32 *Action Units* (AUs). Os AUs por sua vez, representam músculos que são contraídos ou expandidos, e que, quando combinadas, formam as expressões faciais. A Figura 1.1 mostra alguns exemplos de AUs.

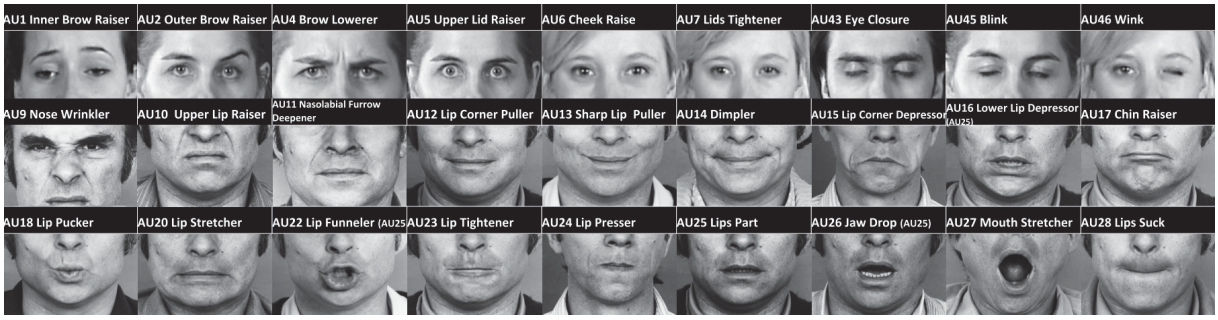


Figura 1.1: Exemplos de AUs, imagem de [Martinez et al., 2017].

A vantagem de utilizar AUs em vez de expressões faciais básicas ou compostas, é a facilidade de modelar as combinações de AUs para atender a necessidades futuras. Conforme mostra a Tabela 1.1, cada expressão facial é composta por uma determinada combinação de AUs,

ficando assim fácil serem adicionadas novas expressões faciais através da modelagem de novas combinações, sem a necessidade de novas anotações ou modificar o método de detecção.

Tabela 1.1: AUs definidos pelo FACS e suas combinações para as expressões faciais básicas (tabela adaptada de [Martinez et al., 2017])

	AUs
FACS	face superior: 1, 2, 4, 5, 6, 7, 43, 45, 46 face inferior: 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20, 22, 23, 24, 25, 26, 27, 28 outro: 21, 31, 38, 39
Raiva	4, 5, 7, 10, 17, 22, 23, 24, 25, 26
Desgosto	9, 10, 16, 17, 25, 26
Medo	1, 2, 4, 5, 20, 25, 26, 27
Felicidade	6, 12, 25
Tristeza	1, 4, 6, 11, 15, 17
Surpresa	1, 2, 5, 26, 27

A detecção de AUs consiste em determinar se um ou mais AUs estão presentes (ativos) na imagem, sendo possível abordá-la como uma classificação binária *single-label*, em que cada AU é processado de forma independente, ou em uma classificação binária *multi-label*, no qual todos são detectados de um vez. Os AUs demonstram co-ocorrência, sendo que alguns AUs, quando ativos, sempre estão acompanhados de outros AUs. A co-ocorrência de AUs pode ser aditiva, na qual a aparência do AU não se altera, ou não-aditiva, na qual a aparência do AU é alterada [Martinez et al., 2017, Benitez-Quiroz et al., 2016]; visando explorar a co-ocorrência dos AUs, a classificação *multi-label* é preferida.

Diversos dos trabalhados focados na detecção de AUs utilizam modelos com extração de características *hand-crafted*. Tais características podem ser baseadas em aparência ou geometria. Características baseadas em aparência utilizam métodos para extrair a textura da face, sendo que os mais utilizados na literatura são: *Histogram of Oriented Objects* (HOG), *Local Binary Patterns* (LBP) e filtros de *Gabor*. Por outro lado, características geométricas são calculadas com base nos pontos fiduciais detectados na face, sendo que podem ser medidos a distância e ângulo entre eles.

Em [Baltrušaitis et al., 2015], os autores utilizaram HOG em conjunto com características geométricas para detectar AUs na base de dados FERA 2015 [Valstar et al., 2015]. No trabalho proposto por [Benitez-Quiroz et al., 2016], foi utilizado a triangulação de *Delaunay*

para definir regiões a serem extraídos os filtros de *Gabor*. O trabalho deles resultou em uma base de dados com mais de um milhão de faces automaticamente anotadas com AUs.

Nos último anos, os esforços na área de análise de expressões faciais vêm sendo direcionados ao reconhecimento de AUs em diferentes poses da cabeça. Conforme a face muda sua posição, seu formato é alterado, resultando até em oclusões em poses extremas, como o caso de pose em perfil. Com o sucesso das redes neurais convolucionais (CNN) em diversas áreas da visão computacional, diversos métodos de reconhecimento de expressões faciais vêm utilizando-as, com resultados superiores aos modelos *hand-crafted*. Normalmente os trabalhos que utilizam CNN focam em processamento holístico, no qual a face é processada como um todo.

No trabalho proposto por [Zoltán et al., 2016] foi utilizada uma CNN simples com apenas três camadas convolucionais para reconhecer AUs em diferentes poses. Os autores realizaram comparativos entre CNN *single-label*, CNN *multi-label* e HOG, sendo que o modelo *single-label* foi o superior. Em [Tang et al., 2017] os autores realizaram o *fine-tuning* da VGG-Face, uma CNN que foi inicialmente desenvolvida para reconhecimento de faces, através de AUs em nove diferentes poses da cabeça. [Li et al., 2017b] propuseram realizar a fusão de uma CNN com características *hand-crafted* para a detecção de AUs em múltiplas poses da cabeça. Por fim, em [Chu et al., 2017] foi desenvolvida uma CNN com cinco camadas convolucionais para predição *multi-label* com uma rede neural recorrente para informações temporais.

Recentemente, foi proposto por [Zhao et al., 2016] um modelo de processamento em *region learning*, no qual a face é divida em 64 regiões fixas dentro de uma CNN, e cada região é processada de forma independente, sendo posteriormente conectadas para a classificação dos AUs de forma conjunta. Além disso, outro trabalho nesta linha foi proposto por [Batista et al., 2017], no qual é utilizado *region learning*, porém com apenas 16 regiões. Ambos os trabalhos apresentaram resultados comparáveis com o estado-da-arte nas bases BP4D [Zhang et al., 2014] e FERA 2017 [Valstar et al., 2017].

O problema de dividir a face em regiões fixas, é o fato de que os AU podem estar localizados no meio das regiões, sendo separados assim em dois ou mais pedaços. Esse problema é mais visível quando se tem regiões menores, como é o caso do modelo proposto por

[Zhao et al., 2016]. Outro fator impactante na separação de regiões é a pose. Com a variação de pose, uma região fixa não acompanha o movimento da face, permanecendo no mesmo lugar. Com isso, regiões treinadas para detectar determinados AUs não conseguem se adaptar a variações de pose, a Figura 1.2 exemplifica o problema.

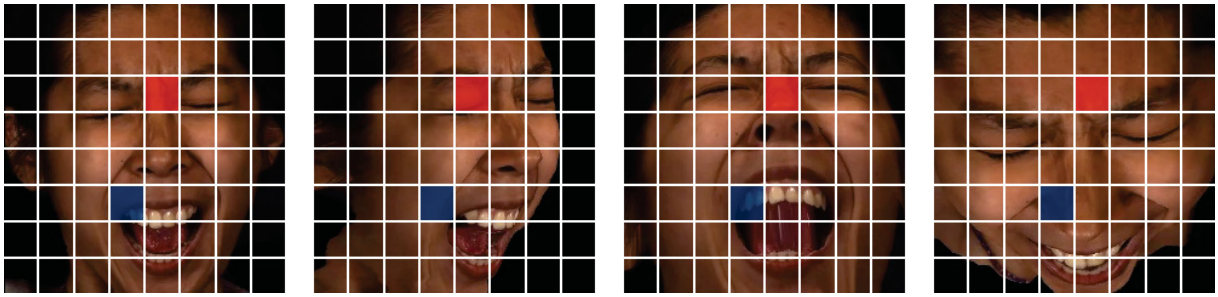


Figura 1.2: Regiões fixas em diferentes poses da cabeça.

[Li et al., 2017a] e [Li et al., 2018] propuseram um modelo de *region learning* usando pontos fiduciais faciais para aprendizado individual. Ambos os métodos fizeram *fine-tuning* da CNN VGG16, exceto que em [Li et al., 2018] é realizado o *fine-tuning* duas vezes, primeiro sem *region learning*, e depois com *region learning* em cima do treino anterior. Em [Li et al., 2017a], o modelo não foi testado em imagens com múltiplas poses, e [Li et al., 2018] demonstraram resultados similares ao método com regiões fixas [Batista et al., 2017]. Além deles, [Ali et al., 2017] propôs um modelo de *region learning* baseado em pontos fiduciais para detecção em múltiplas poses. No modelo proposto pelos autores, é realizada apenas uma convolução antes da camada de regiões, e dentro de cada região existem 5 camadas convolucionais, que posteriormente são concatenadas para predição *multi-label*.

Além da detecção de AUs baseada em características estáticas, a detecção pode explorar a evolução dos AUs durante o tempo, através de características temporais. Informações temporais são muito importantes para a detecção de AUs, pois as expressões faciais acontecem por um período mais longo que apenas um frame, então, utilizar as informações dinâmicas de um vídeo ajuda os métodos a melhorares suas predições. Funcionando em *loop*, as redes neurais recorrentes (RNN) são uma boa forma de realizar a modelagem temporal, .

Conforme exemplifica a Figura 1.3, as RNNs funcionam em uma estrutura de repetição, na qual o estado anterior da rede é usado para ajudar na predição atual, fornecendo, assim, informações temporais. Um grande problema das RNNs é o chamado *vanishing gradient*, que

nada mais é que com o passar do tempo (iterações do *loop*), os pesos da rede neural vão ficando tão pequenos que desaparecem. Para contornar esse problema, foi proposta uma rede neural recorrente chamada *Long-Short Term Memory* (LSTM) [Hochreiter e Schmidhuber, 1997]. Diferente das RNNs simples, que possuem apenas um *gate*, a LSTM possui quatro, possibilitando que a rede consiga escolher o que deve ser removido, atualizado, ou adicionado e, desta forma, contornar o problema do *vanishing gradient*.

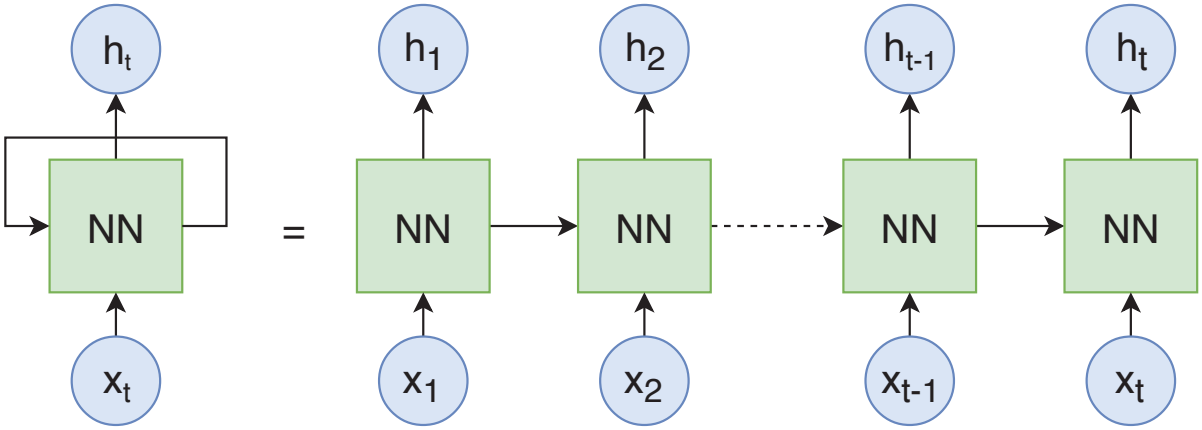


Figura 1.3: Estrutura de repetição de uma rede neural recorrente.

No trabalho proposto por [Li et al., 2017a], os autores adicionaram ao final da CNN uma camada LSTM, transformando-a em uma LRCN (*Long-term Recurrent Network*), possibilitando assim, que a rede seja treinada *end-to-end*. Com a adição da LSTM, os autores conseguiram um melhoramento significativo em comparação ao modelo estático proposto por eles. [Chu et al., 2017] propôs um método semelhante, porém os autores conectaram uma camada LSTM no final da CNN, sendo treinada separadamente. Posteriormente é feita a fusão da saída da LSTM com a saída da CNN, resultando nas predições. Ambos os trabalhos foram desenvolvidos somente com imagens frontais. Já no modelo desenvolvido por [He et al., 2017], foram treinadas 90 redes neurais convolucionais, uma para cada pose (9 poses) e AU (10 AUs). Cada CNN é posteriormente ligada a uma rede LSTM bidirecional para auxílio na predição dos AUs. O modelo desenvolvido com modelagem temporal, porém, obteve resultados semelhantes aos modelos estáticos.

Neste contexto, este trabalho propõe regiões dinâmicas de aprendizado (DRL) utilizando pontos fiduciais faciais para criar regiões para aprendizado individual e modelagem temporal. O modelo foi desenvolvido utilizando a rede VGG-Face [Parkhi et al., 2015] e a rede neural

recorrente LSTM [Hochreiter e Schmidhuber, 1997]. Para avaliar a habilidade do trabalho desenvolvido em se adaptar às diferentes poses, a base de dados FERA 2017 [Valstar et al., 2017] foi utilizada, sendo que é a única disponível com foco em detecção de AUs em múltiplas poses da cabeça. Vale ressaltar que este trabalho resultou nas seguintes publicações: [Batista et al., 2017], [Zavan et al., 2017] e [Albiero et al., 2018].

O restante deste trabalho está dividido da seguinte maneira: o Capítulo 2 apresenta a detecção da face, detecção dos pontos fiduciais faciais, e a implementação das regiões dinâmicas de aprendizado; o Capítulo 3 demonstra a extração de características, e a arquitetura do modelo temporal proposto; os resultados experimentais são exibidos no Capítulo 4; e, por fim, o Capítulo 5 apresenta a conclusão do trabalho.

2 Região dinâmica de aprendizado (DRL)

Neste capítulo são abordados os conceitos envolvidos no desenvolvimento das regiões dinâmicas de aprendizado para análise de expressões faciais através da detecção de AUs. O capítulo esta organizado da seguinte forma: a Seção 2.1 descreve os procedimentos utilizados para detectar as faces; a Seção 2.2 demonstra a detecção e seleção dos pontos fiduciais faciais; e, por último, a Seção 2.3 descreve a implementação da região dinâmica de aprendizado.

2.1 Detecção da face

Para detectar as faces, foram utilizados o detector de objetos Faster R-CNN [Ren et al., 2015], juntamente com a rede neural convolucional VGG16 [Simonyan e Zisserman, 2015]. Apesar de ser um detector genérico de objetos, a Faster R-CNN foi re-treinada para detecção de faces na base de dados WIDER FACE [Yang et al., 2016]. O treino foi iniciado utilizando pesos pré-treinados na base de dados ImageNet [Simon et al., 2016], que consiste em mil classes de objetos, permitindo assim iniciar o treinamento com filtros robustos. Após o treino, o modelo foi validado no *benchmark* FDDB [Jain e Learned-Miller, 2010], com o qual foram obtidos resultados comparáveis ao estado da arte.

2.2 Detecção e seleção dos pontos fiduciais faciais

Por lidar com múltiplas poses da cabeça, detectores simples como os presentes na biblioteca Dlib não foram capazes de encontrar os pontos fiduciais faciais em todas as poses. Foi então utilizado o detector desenvolvido por [Bulat e Tzimiropoulos, 2017], que é baseado

em redes neurais convolucionais. Atualmente ele é o detector mais preciso disponível, sendo capaz de detectar pontos em coordenadas 2D e 3D. Para evitar predições erradas, informações temporais foram utilizadas para suavizar as detecções, sendo que a mediana dos pontos fiduciais de 10 *frames* foram utilizados para iniciar as detecções. A cada novo *frame*, os pontos fiduciais encontrados foram comparados com a mediana anterior, e caso a distância euclidiana fosse maior que 25 pixels, o ponto anterior era utilizado no lugar. A Figura 2.1 mostra um exemplo da melhoria nas detecções através da suavização temporal em diferentes poses da cabeça.

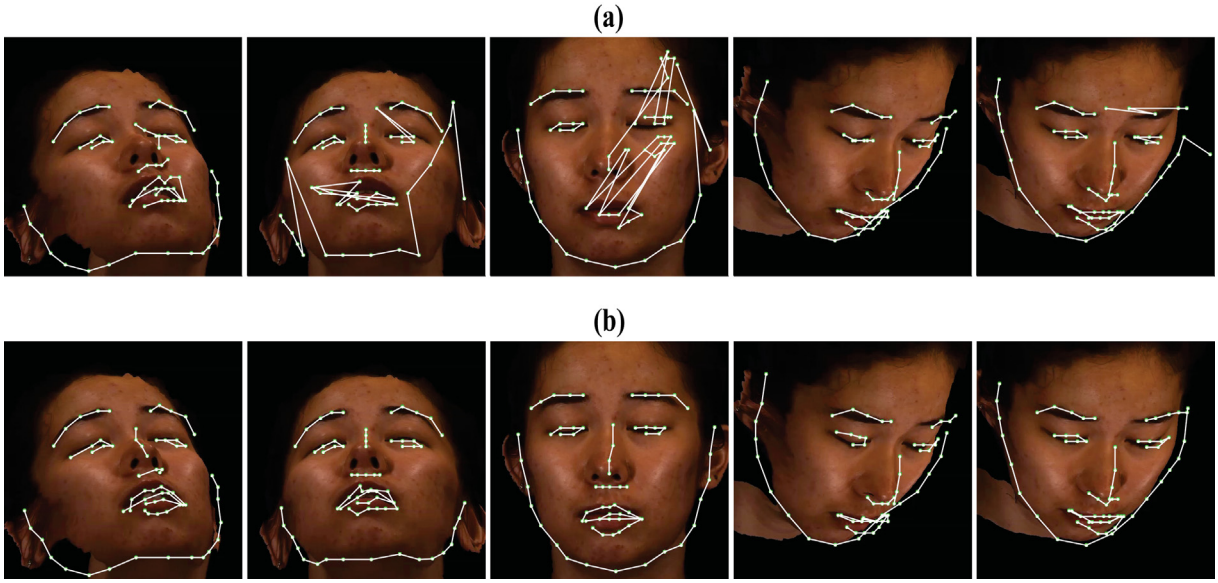


Figura 2.1: Comparação entre a detecção de pontos fiduciais sem suavização temporal (a) e com (b) para diferentes poses da cabeça.

Para criar os pontos fiduciais faciais a serem usados pela região dinâmica de aprendizado, 20 pontos foram selecionados no centro da localização de cada AU. Exceto para o AU01, AU06 e AU17, todos os pontos selecionados estão situados exatamente no local obtido pelas detecções. Para o AU06 e AU17, os pontos foram deslocados em direção ao centro da localização do AU, que são, respectivamente, o meio da bochecha e do queixo. Para o AU01, os pontos fiduciais na parte internar da sobrancelha foram usados em conjunto com o ponto da parte externa, o que ajuda a detectar se a sobrancelha interna foi levantada ou não. A Figura 2.2 mostra os pontos fiduciais faciais utilizados.

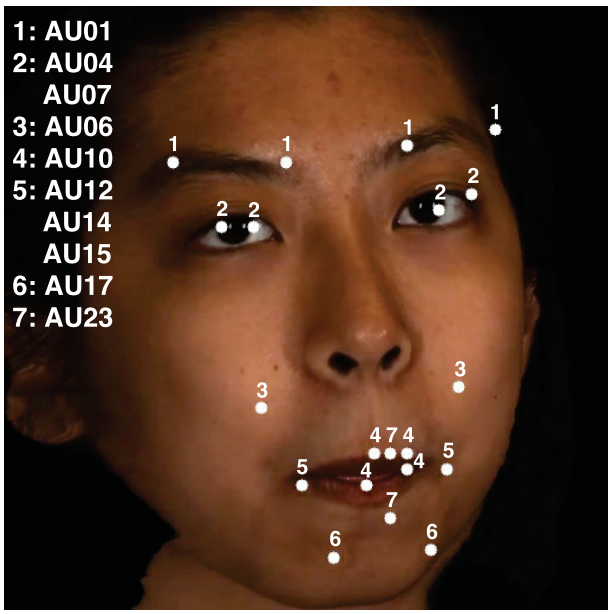


Figura 2.2: Pontos fiduciais faciais utilizados e seus *Action Units* correspondentes.

2.3 Arquitetura da região dinâmica de aprendizado

Para implementar a região dinâmica de aprendizado, a CNN VGG-Face [Parkhi et al., 2015] foi escolhida. A VGG-Face é uma rede neural convolucional baseada na rede VGG16 [Simonyan e Zisserman, 2015]. A VGG-Face foi treinada em milhões de imagens para reconhecimento facial, resultando então em uma rede *expert* em faces. Ao utilizar *transfer-learning* do conhecimento de extração de características faciais da VGG-Face, a DRL inicia o treino com filtros mais robustos, comparado com a inicialização randômica ou inicialização com pesos da ImageNet [Simon et al., 2016].

A região dinâmica de aprendizado é aplicada após a última camada convolucional da VGG-Face, que contém 512 filtros. A DRL comprime estes 512 filtros em 20 regiões contendo 16 filtros cada, agindo como um *embedding*. A entrada do modelo proposto é composto por uma imagem RGB de 224x224 pixels sem alinhamento, e 20 pontos fiduciais faciais. A imagem é então primeiramente processada de forma holística pela VGG-Face e, então, a DRL usa os pontos fiduciais para recortar 20 regiões de 3x3 pixels cada. As regiões são aumentadas para 6x6 pixels e filtradas por uma camada convolucional de 3x3 individualmente, gerando uma região final de 4x4 pixels. Finalmente, cada região é conectada a uma camada localmente conectada, e então todas são concatenadas e enviadas a duas camadas totalmente conectadas

para a predição *multi-label* de 10 AUs. Antes de cada camada totalmente conectada, é aplicado *dropout* [Srivastava et al., 2014] com probabilidade de 50% para evitar *overfitting*. A Figura 2.3 exemplifica o modelo proposto.

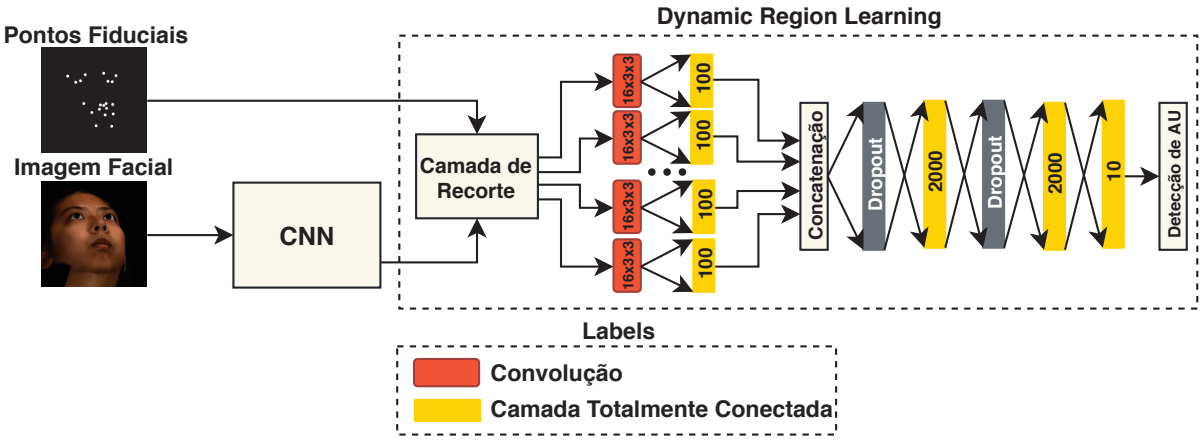


Figura 2.3: Exemplo do modelo proposto. A face primeiramente é processada pela rede VGG-Face. Em seguida, utilizando pontos fiduciais faciais, a face processada é recortada em 20 regiões, que são filtradas individualmente e concatenadas para predição *multi-label* de AUs.

3 Modelagem Temporal

Neste capítulo são abordados os conceitos envolvidos na modelagem temporal proposta. O capítulo está organizado da seguinte forma: a Seção 3.1 descreve a criação inicial dos vetores de características; e a Seção 3.2 descreve a implementação da arquitetura da rede neural recorrente proposta.

3.1 Extração inicial de vetores de características

Para criar os vetores de características de entrada para a rede neural recorrente, foi utilizada a saída da camada anterior à saída dos AUs da DRL, contendo 2000 características. O processo de extração consiste em executar a DRL para todas as imagens, da base de treino e validação, e salvar um vetor de características que representa cada imagem. Os vetores são agrupados por vídeo, gerando assim uma matriz de características que represente cada um dos vídeos da base de dados.

3.2 Arquitetura da rede neural recorrente

Utilizando os vetores de características extraídos pela DRL, foi implementada uma camada LSTM [Hochreiter e Schmidhuber, 1997] com 500 neurônios. Após ela, uma camada totalmente conectada com também 500 neurônios é aplicada, e finalmente a saída são 10 AUs. Entre as camadas finais, e dentro da camada LSTM, é aplicado *dropout* [Srivastava et al., 2014] com probabilidade de 50% para reduzir *overfitting*

Como mostra a Figura 3.1, utilizando t vetores de características, são instanciados t LSTMs, no qual o estado temporal de cada vetor de características é enviado para o vetor

seguinte. Logo após, as camadas seguintes são conectadas a cada uma das células LSTMs e concatenadas para predição de t x 10 AUs. O estado temporal final da sequência não é levado para sequências futuras, e o erro da predição em comparação com o *ground truth* é calculado para todos as t predições de forma simultânea. Desta forma, é possível modelar a evolução temporal de cada vídeo separadamente, sem que o estado das características de vídeos anteriores influenciem no estado do vídeo atual.

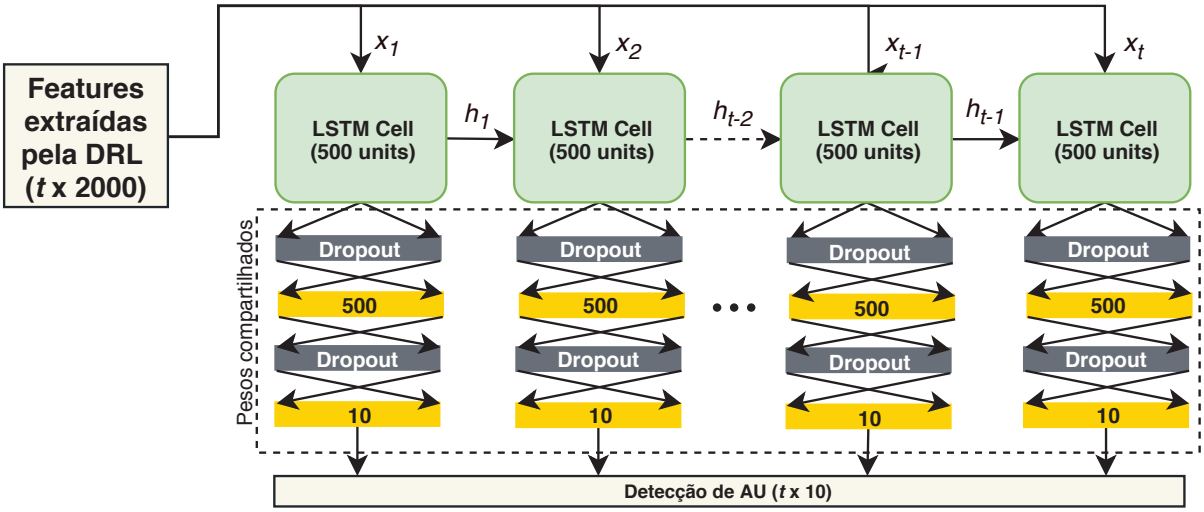


Figura 3.1: Modelo proposto de LSTM. Neste modelo é utilizado uma sequência de t vetores de características (t representando tempo), sendo que o estado temporal não é levado para a próxima sequência.

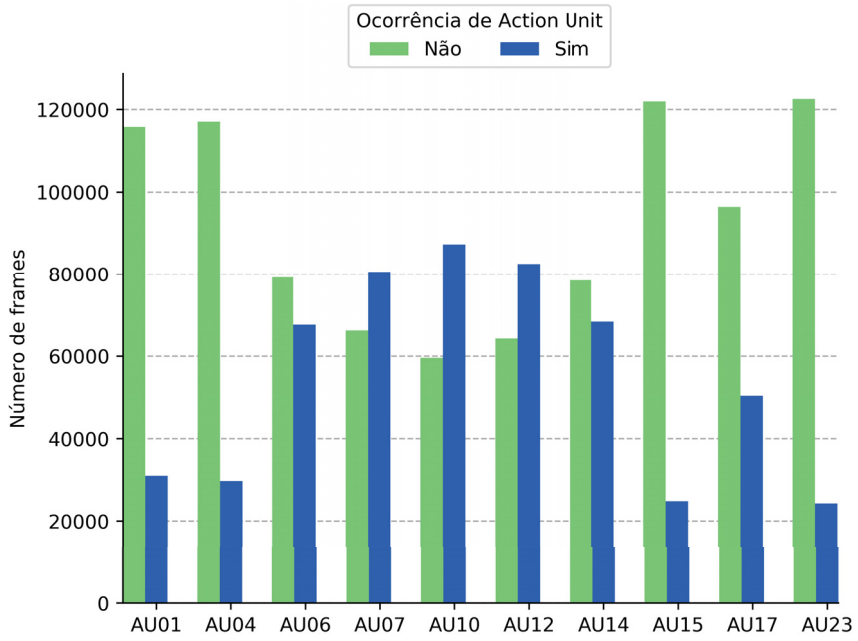
4 Resultados experimentais

4.1 Base de dados

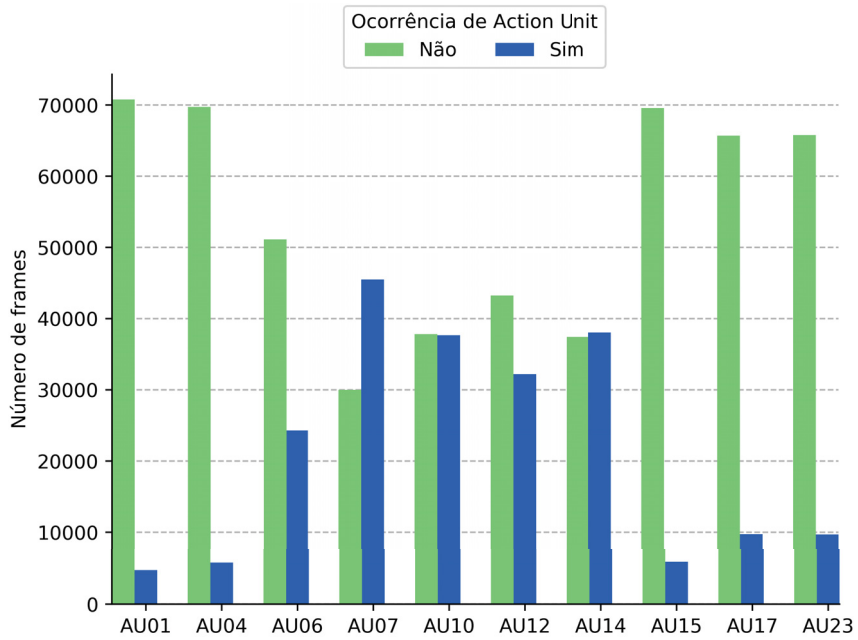
Para os experimentos, foi utilizada a base de dados Facial Expression Recognition and Analysis Challenge (FERA 2017) [Valstar et al., 2017], sendo a única base de AUs focada em múltiplas poses da cabeça. A base de dados é composta por vídeos 2D renderizados em nove poses diferentes da base de dados BP4D-Spontaneous [Zhang et al., 2014]. Exemplos das poses podem ser vistos na Figura 4.1. A base contém anotações de dez AUs, sendo eles: AU01 (*inner brow raiser*); AU04 (*brow lowerer*); AU06 (*cheek raiser*); AU07 (*lid tightener*); AU10 (*upper lip raiser*); AU12 (*lip corner puller*); AU14 (*dimpler*); AU15 (*lip corner depressor*); AU17 (*chin raiser*); and AU23 (*lip tightener*). Além disso, a base de dados é divida em três partes: treino, validação e teste. Por enquanto, somente o treino e a validação estão disponíveis publicamente.

Figura 4.1: Exemplo das 9 poses na base de dados FERA17.

A base de dados contém muito mais exemplos negativos do que positivos e, conforme mostra a Figura 4.2, ambas as partes de treino e validação sofrem com desbalanceamento, porém, o desbalanceamento é ainda mais acentuado na parte de validação, o que faz com que modelos treinados na parte de treino sofram mais para detectar certos AUs na base de validação.



(a) Desbalanceamento de classes na base de treino.



(b) Desbalanceamento de classes na base de validação.

Figura 4.2: Desbalanceamento de classes na base de dados FERA 2017

Devido a base de dados conter muito mais exemplos negativos do que positivos, a métrica padrão de acurácia não consegue expressar bem a performance do modelo. Por isso foi utilizada a métrica F_1-score. O F_1-score não leva em consideração os casos negativos, apenas os falso-positivos, positivos e falso-negativos, sendo assim mais robusta ao desbalanceamento de classes presente da base.

4.2 Rede Neural Convolucional

4.2.1 Treinamento CNN

No treinamento do modelo proposto foi utilizado o *framework* Keras com *backend* TensorFlow. Como a base de dados contém mais de um milhão de *frames*, e *frames* consecutivos contêm informações redundantes, somente 10% da base de dados foram usadas para treino, totalizando 133.458 *frames* separados ao selecionar o primeiro a cada dez *frames*. Para otimização, foi usado *Stochastic Gradient Descent* (SGD), iniciando com *learning rate* de 0,001, e *mini-batches* de 64 imagens. Depois de cada *epoch* (quando o treino visualizou todas as imagens), 10% da base de validação são utilizadas para avaliar o modelo, e caso a *loss* não reduza por duas *epochs* consecutivas, o *learning rate* é reduzido. Finalmente, para *early stop* é utilizado a métrica F_1-score, em que caso o modelo não melhore por cinco *epochs* consecutivas, o treino é finalizado. Todos os resultados mostrados nas próximas seções foram obtidos testando com a base de validação completa (681.462 *frames*).

4.2.2 DRL vs. VGG-Face com *fine-tuning*

Para avaliar a performance da região dinâmica de aprendizado, primeiro, a rede VGG-Face foi retreinada com *fine-tuning* na base de dados sem mudanças na sua arquitetura. Após, a DRL foi implementada conforme descrito na Seção 2.3. Conforme mostra a Figura 4.3, a região dinâmica de aprendizado teve performance melhor que a VGG-Face para quase todos os AUs, menos o AU07. Para os outros AUs, o AUAU01, AU04 e AU06 tiveram o maior aumento no resultado, pelo motivo de que esses pontos fiduciais faciais para esses AUs são mais fáceis de localizar em diferentes poses. No geral, a média do F_1-score da rede VGG-Face com *fine-tuning* é de 0,545, e do modelo proposto é 0,582, alcançando uma melhora de 6,79%.

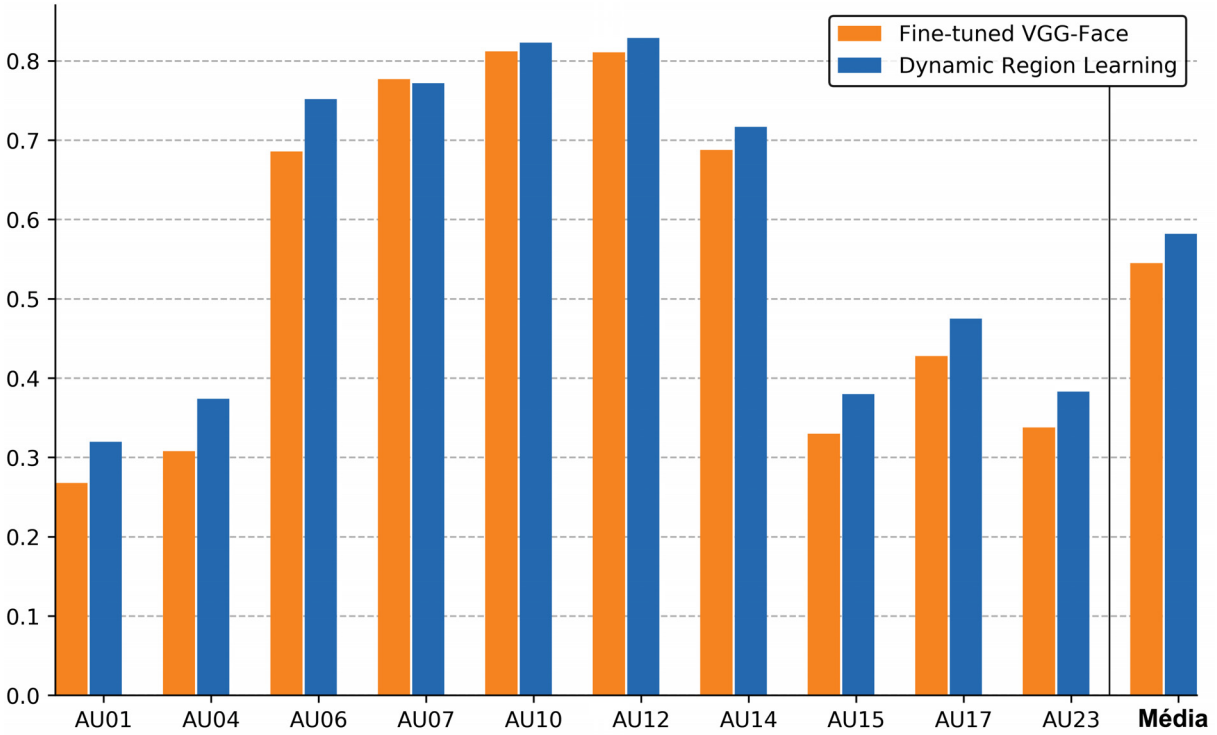


Figura 4.3: Resultados da VGG-Face com e sem DRL, calculados usando F_1-score para detecção de AUs.

4.2.3 Análise por pose da cabeça

Os resultados obtidos para cada uma das nove poses são apresentados na Tabela 4.1. A pose da cabeça que teve melhor performance não foi a frontal, mas sim a pose com uma pequena rotação horizontal (pose 5). As poses que também contêm rotação horizontal (pose 4 e 2) mostram resultados semelhantes à pose 5. Além disso, a pose frontal teve resultados piores que várias poses com rotação horizontal, ficando em quinto lugar no resultado médio. Então, pode-se presumir que a rotação no eixo horizontal ajuda a DRL a detectar melhor os AUs. As poses rotacionadas para baixo (pose 7, 8 e 9) mostraram a pior performance. Quando a face é rotacionada para baixo, os AUs inferiores são menos visíveis, e também, os AUs superiores são afetados pela estrutura da face. A pose 9 foi a mais difícil, sendo que juntamente com a rotação para baixo, ela não possui rotação horizontal. Comparado ao FERA 2017 [Valstar et al., 2017], cujo *baseline* teve um resultado $max - min$ F_1-score de 0,250, o modelo proposto demonstra boa adaptação às nove poses, com um resultado $max - min$ F_1-score de 0,079, mostrando uma melhora de 68,4%.

Tabela 4.1: Resultados da DRL VGG-Face calculados com F_1-score para detecção de AU nas nove poses da base de dados (melhores resultados por AU em verde e piores em vermelho).

AUs	Poses								
	1	2	3	4	5	6	7	8	9
01	**0,372**	0,333	0,289	0,365	0,339	0,309	0,308	0,286	0,307
04	0,376	0,393	0,324	0,402	**0,433**	0,397	0,332	0,364	0,364
06	0,771	**0,784**	0,768	0,750	0,763	0,754	0,737	0,712	0,727
07	0,775	0,787	0,781	0,787	**0,802**	0,787	0,764	0,749	0,707
10	0,826	**0,830**	**0,830**	0,823	0,833	0,827	0,811	0,812	0,817
12	0,827	0,838	0,831	0,837	**0,851**	0,846	0,814	0,807	0,809
14	0,730	**0,749**	0,727	0,743	0,726	0,711	0,703	0,694	0,672
15	0,387	0,391	0,364	0,413	0,406	**0,415**	0,367	0,382	0,286
17	0,485	0,486	0,461	0,519	**0,526**	0,493	0,482	0,475	0,365
23	0,419	0,432	**0,437**	0,395	0,419	0,401	0,336	0,325	0,251
Média	0,597	0,602	0,581	0,603	**0,610**	0,594	0,565	0,560	0,531

4.2.4 Aplicação da DRL em outras CNNs

A fim de verificar a habilidade da DRL em melhorar o desempenho de redes neurais convolucionais em geral, foram realizados testes com algumas CNNs populares. Para tal, foram utilizadas as redes: ResNet [He et al., 2016]; Inception [Szegedy et al., 2016]; e Xception [Chollet, 2017]. O treino destas CNNs seguiu as configurações previamente descritas, e foram iniciadas realizando *transfer-learnig* com pesos da ImageNet [Simon et al., 2016]. Para cada CNN foram realizados dois treinos: um sem DRL e outro com. No treino sem DRL, foi realizado o *fine-tuning* de cada uma das redes com a sua arquitetura original. Já no treino com DRL, foi utilizada a saída da última camada convolucional de cada rede, e aplicada a DRL sem alterações, conforme descrita na Seção 2.3.

A CNN ResNet [He et al., 2016] foi criada com o conceito de camadas residuais, no qual camada anteriores são adicionadas às saídas da camadas seguintes. Conforme mostra a Figura 4.4, a ResNet com DRL obteve resultados melhores para seis AUs, resultados semelhantes para dois AUs, e pior para o AU01, que ficou bem abaixo da ResNet sem DRL. Por outro lado, o AU15, AU17 e AU23 obtiveram resultados superiores bem significativos. Por último, a média do F_1-score da rede ResNet sem DRL foi de 0,521, e da ResNet com DRL de 0,549, obtendo uma melhora de 5,37%.

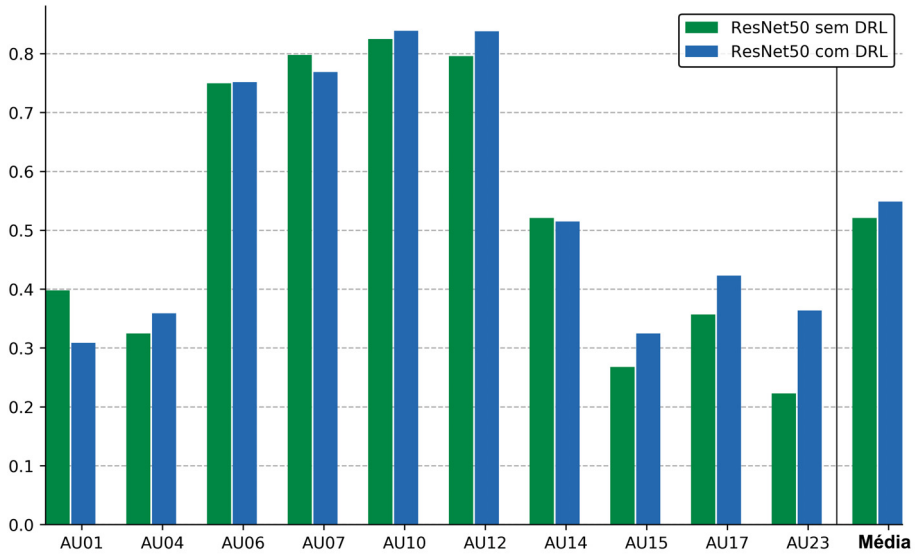


Figura 4.4: Resultados da ResNet50 com e sem DRL, usando F_1-score para detecção de AUs.

A rede Inception [Szegedy et al., 2016] foi desenvolvida pela Google com o objetivo de acelerar o processo de treino. Ela é montada em cima de blocos que utilizam convoluções com diferentes quantidades de filtros para reduzir a quantidade de parâmetros, permitindo assim a rede ser mais profunda. Conforme mostra a Figura 4.5, a Inception com DRL obteve resultados superiores para sete Action Units, sendo que o AU14 obteve um resultado mais expressivo, com uma melhora de 31,49%. Na média geral, o F_1-score teve um aumento de 4,83%, com uma média de 0,538 e 0,564 para a Incetpion sem e com DRL, respectivamente.

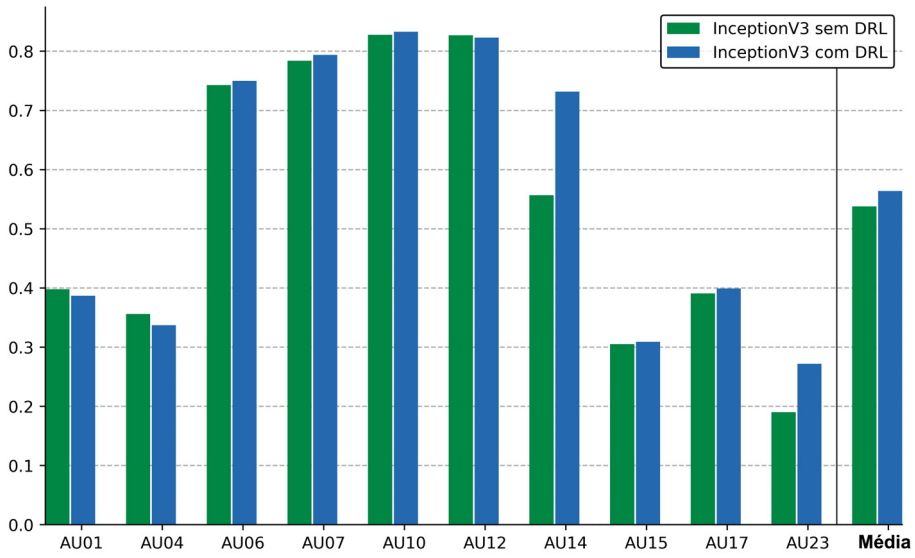


Figura 4.5: Resultados da InceptionV3 com e sem DRL, calculados usando F_1-score para detecção de AUs.

A Xception [Chollet, 2017] é uma rede neural convolucional desenvolvida pelo autor do *framework* Keras. Ela foi criada com o objetivo de ser uma releitura do módulo *inception* da rede Inception [Szegedy et al., 2016], no qual as convoluções dentro de cada camada *inception* são realizadas de forma mais eficiente, melhorando assim o desempenho da rede. Conforme mostra a Figura 4.6, a adição da DRL na Xception não surgiu o mesmo efeito que nas outras redes, sendo que para a maioria dos AUs, o resultado foi igual. Já para o AU04 e AU07, a adição da DRL diminuiu os resultados. Por outro lado, semelhante ao que ocorreu na rede Inception, o AU14 foi obteve uma melhora significativa de 18,84%. Por fim, a média do F_1-score foi de 0,555 para a Xception sem DRL e de 0,559 para a Xception com DRL.

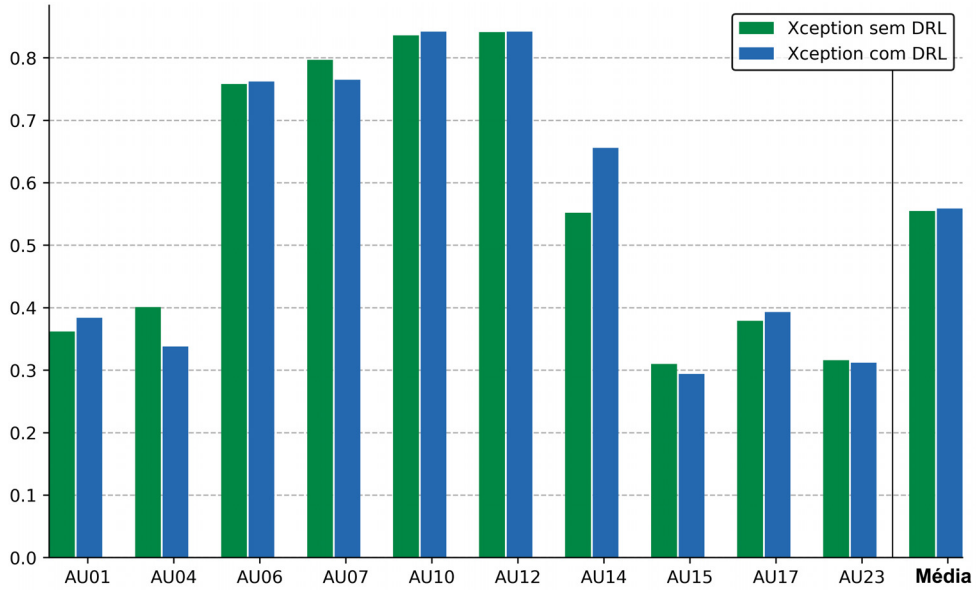


Figura 4.6: Resultados da Xception com e sem DRL, usando F_1-score para detecção de AUs.

Por fim, a Tabela 4.2 mostra um comparativo entre as três CNNs e a VGG-Face com aplicação da DRL. Como pode-se observar, a DRL VGG-Face obteve resultados melhores para cinco AUs, além de um F_1-score médio bem superior, o que pode-se atribuir ao fato da utilização de pesos treinados em milhões de faces, diferente das outras redes que realizaram *fine-tuning* com pesos treinados na ImageNet [Simon et al., 2016]. Vale ressaltar que a rede DRL Inception obteve resultados melhores para três AUs, e a DRL Xception para dois.

Tabela 4.2: Comparativo entre CNNs com DRL para detecção de AUs usando F_1-score (melhores resultados por AU em verde e piores em vermelho).

Métodos	AUs										
	01	04	06	07	10	12	14	15	17	23	Média
DRL ResNet50	0,309	0,359	0,752	0,769	0,839	0,838	0,515	0,325	0,423	0,364	0,549
DRL InceptionV3	**0,387**	0,337	0,750	**0,794**	0,833	0,823	**0,732**	0,309	0,399	0,272	0,564
DRL Xception	0,384	0,338	0,762	0,768	**0,842**	**0,842**	0,656	0,294	0,393	0,312	0,559
DRL VGG-Face	0,320	**0,374**	**0,752**	0,772	0,823	0,829	0,717	**0,380**	**0,475**	**0,383**	**0,582**

4.3 Rede Neural Recorrente

4.3.1 Treinamento RNN

No treinamento da rede neural recorrente, foi utilizado o *framework* Keras com *backend* TensorFlow. Diferente do treino das CNNs, foi utilizado a base de treino inteira para treinamento. Para otimização, o otimizador RMSProp foi escolhido, iniciando com *learning rate* de 0,001. O treino foi executado com *mini-batches* de 64 sequências contendo 700 vetores de característica cada. Como a maioria dos vídeos possuem tamanho menor que 700 *frames*, foram definidas sequências de 700 vetores de características para treino, o que possibilita que cada vídeo seja inserido de forma completa dentro de uma sequência. Para preencher as sequências com vídeos menores que 700 *frames*, é realizado um *padding* com 9s ao final do vetor, o qual é ignorado pelo modelo. Ao realizar os testes na base de validação, não é necessário realizar *padding*, pois como é testado vídeo a vídeo, é possível utilizar o tamanho real do vídeo como entrada ao modelo.

4.3.2 Comparação entre modelo estático e temporal

A Figura 4.7 mostra os resultados obtidos pelo modelo com LSTM em comparação a DRL VGG-Face. Como pode-se observar, a modelagem temporal obteve melhoras significativas para o AU01 e AU14, e uma pequena melhora nos AU07 e AU12. Já para o AU15 e AU23, os resultados foram semelhantes. Além disso, o modelo obteve resultados piores para o AU06, AU04, AU10 e AU17. Uma explicação para a performance ter sido pior em alguns AUs, e melhor em outros, é a possibilidade da rede ter sofrido com *overfitting*. Como a DRL VGG-Face foi

treinada na base de treino, e posteriormente a LSTM foi treinada na mesma base, características correspondentes a alguns AUs podem ter sido aprendidas demais, resultando em uma queda na performance. Por fim, o F_1-score médio do modelo com LSTM foi de 0,593, ficando 1,89% acima da DRL VGG-Face.

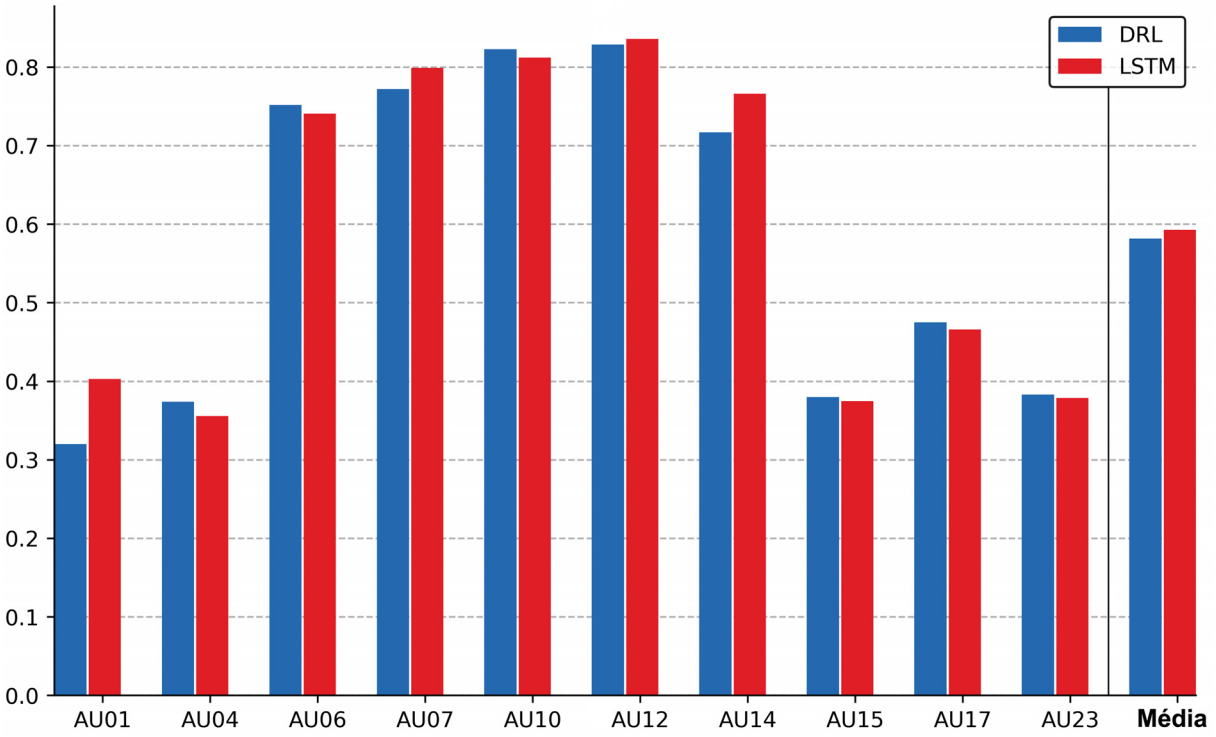


Figura 4.7: Resultados da DRL VGG-Face e LSTM, calculados usando F_1-score para detecção de AUs.

4.3.3 Análise por pose da cabeça

A Tabela 4.3 mostra os resultados obtidos para cada uma das nove poses. Assim como no método sem modelagem temporal, a pose da cabeça com melhor performance foi a pose com uma pequena rotação horizontal (pose 5). Diferente do modelo estático, a pose que teve o segundo melhor resultado, foi a pose com pequena rotação horizontal e rotação vertical para cima (pose 1). Em terceiro lugar, a pose frontal (pose 6) e com maior rotação horizontal e rotação para cima (pose 2). Da mesma maneira que na DRL VGG-Face, as poses com pior resultado foram as poses rotacionadas para baixo (pose 7, 8 e 9).

As melhorias por pose em comparação ao modelo estático foram de: Pose 1, 2,51%; Pose 2, 1,00%; Pose 3, 1,89%; Pose 4, 0,50%; Pose 5, 1,64%; Pose 6, 2,36%; Pose 7, 3,36%; Pose

8, 1,79%; Pose 9, 2,82%; e na média geral, 1,89%. Além disso, comparado a média $max - min$ F_1-score de 0,079, o modelo com LSTM, obteve um melhoramento de 6,33%, obtendo um resultado $max - min$ F_1-score de 0,074.

Tabela 4.3: Resultados do modelo temporal calculados com F_1-score para detecção de AU nas nove poses da base de dados (melhores resultados por AU em verde e piores em vermelho).

AUs	Poses								
	1	2	3	4	5	6	7	8	9
01	**0,456**	0,407	0,366	0,423	0,420	0,383	0,444	0,354	0,411
04	0,368	0,363	0,326	0,384	**0,418**	0,366	0,313	0,350	0,332
06	0,756	**0,763**	0,745	0,733	0,760	0,745	0,726	0,712	0,725
07	0,817	**0,829**	0,813	0,804	0,815	0,807	0,780	0,771	0,753
10	0,819	**0,822**	0,812	0,812	0,814	0,819	0,802	0,801	0,804
12	0,845	0,842	0,832	0,859	**0,863**	0,852	0,816	0,808	0,807
14	0,776	0,770	0,772	0,760	**0,777**	0,775	0,773	0,744	0,747
15	0,373	0,388	0,355	0,401	0,413	**0,437**	0,354	0,377	0,288
17	0,487	0,478	0,456	0,499	**0,515**	0,484	0,467	0,464	0,360
23	0,428	0,418	**0,438**	0,384	0,403	0,411	0,361	0,322	0,236
Média	0,612	0,608	0,592	0,606	**0,620**	0,608	0,584	0,570	0,546

4.4 Comparação com outros métodos da literatura

Como mostra a Tabela 4.4, comparado ao método base [Valstar et al., 2017] e o modelo com CNN e características *hand-crafted* [Li et al., 2017b], a DRL VGG-Face teve melhor performance para todos os AUs, com aumentos significativos. Em comparação ao modelo com regiões fixas [Batista et al., 2017], a região dinâmica de aprendizado teve resultados melhores para quase todos os AUs, menos para os AU01 e AU07. A possível razão é que esses AUs ficaram no centro das regiões fixas, ao contrário dos demais.

O método com regiões de aprendizado usando pontos fiduciais faciais [Li et al., 2018], teve performance inferior ao método proposto, com resultado maior apenas para o AU07. A conclusão é que ele não comprimiu os filtros na camada de região, e também realizou *fine-tuning* duas vezes, o que pode ter ocasionado em *overfitting*. O método [Tang et al., 2017] alcançou resultado melhor para quatro AUs: AU07, AU10, AU12 e AU23. Para os outros seis AUs, a DRL VGG-Face teve melhor performance. Juntamente com a melhora na média, o modelo proposto é

multi-label com uma única rede, tornando o tempo de predição e treino mais rápido que o modelo [Tang et al., 2017], que é *single label*. Finalmente, a região dinâmica de aprendizado alcançou resultados melhores para mais AUs que todos os outros métodos, superando o estado-da-arte com um F_1-score médio de 0,582.

Tabela 4.4: Comparativo com outros métodos estáticos para detecção de AUs usando F_1-score (melhores resultados por AU em verde e piores em vermelho).

Métodos	AUs										
	01	04	06	07	10	12	14	15	17	23	Média
[Valstar et al., 2017]	0,154	0,172	0,564	0,727	0,692	0,647	0,622	0,146	0,224	0,207	0,416
[Li et al., 2017b]	0,288	0,225	0,600	0,749	0,751	0,730	0,606	0,246	0,284	0,248	0,473
[Batista et al., 2017]	**0,345**	0,278	0,677	0,794	0,785	0,762	0,692	0,267	0,364	0,250	0,521
[Li et al., 2018]	0,272	0,332	0,699	**0,808**	0,834	0,802	0,621	0,251	0,342	0,261	0,522
[Tang et al., 2017]	0,304	0,362	0,712	0,779	**0,836**	**0,840**	0,697	0,353	0,442	**0,475**	0,580
DRL VGG-Face	0,320	**0,374**	**0,752**	0,772	0,823	0,829	**0,717**	**0,380**	**0,475**	0,383	**0,582**

A Tabela 4.5 mostra a comparação com o único modelo na literatura desenvolvido com informações temporais que reportou resultados na base de dados FERA 2017. Em comparação ao método [He et al., 2017], que utiliza uma rede LSTM bidirecional, o modelo temporal desenvolvido utilizando uma rede LSTM comum, atingiu resultados superiores para todos os AUs, com um resultado médio superior de 13,60%.

Tabela 4.5: Comparativo com outros métodos temporais para detecção de AUs usando F_1-score (melhores resultados por AU em verde e piores em vermelho).

Métodos	AUs										
	01	04	06	07	10	12	14	15	17	23	Média
[He et al., 2017]	0,369	0,264	0,678	0,763	0,801	0,796	0,664	0,269	0,366	0,248	0,522
DRL + LSTM	**0,403**	**0,356**	**0,741**	**0,799**	**0,812**	**0,836**	**0,766**	**0,375**	**0,466**	**0,379**	**0,593**

4.5 Discussão

Após a validação experimental, ressaltam-se alguns detalhes sobre os experimentos: as redes neurais convolucionais treinadas para validar a capacidade da DRL (três primeiras linhas da Tabela 4.2), foram capazes de obter resultados superiores a maioria dos métodos disponíveis

na literatura (Tabela 4.4), ficando somente atrás do método proposto por [Tang et al., 2017]; a DRL foi capaz de se adaptar bem às poses da base, confirmando a importância do uso de regiões dinâmicas através de pontos fiduciais faciais, ao invés de regiões fixas. Os resultados mostraram que, diferente do esperado, a pose com o melhor resultado não foi a frontal, mais sim a pose com uma pequena rotação horizontal. Foi também verificado que rotações horizontais ajudam na detecção dos AUs, e que rotações verticais para baixo prejudicam a detecção. A adição da DRL nas CNNs resultou em melhoramentos significativos para três das quatro redes testadas, mostrando que não foi por acaso a melhora obtida com a aplicação na VGG-Face. Por outro lado, a junção do modelo estático com a modelagem temporal em forma de redes LSTM melhorou os resultados, diminuindo a diferença entre a pose com melhor e pior F_1-score.

5 Conclusão

Este trabalho apresentou regiões dinâmicas de aprendizado, redes neurais convolucionais e redes neurais recorrentes para detecção *multi-label* de AUs em imagens com múltiplas poses da cabeça. Os experimentos comprovaram que a aplicação de regiões dinâmicas de aprendizado tem um resultado melhor comparado com outros métodos da literatura, inclusive sob as regiões fixas, sendo capaz de adaptar-se as modificações da pose e focar no centro dos AUs.

Vale ressaltar que além do aumento de desempenho de 6,79% da DRL na VGG-Face, a aplicação da DRL em outras CNNs populares resultou em melhoras significativas em duas das três redes testadas: na ResNet, o melhoramento foi de 5,37%; e na rede Inception, 4,83%.

Além disso, utilizando vetores de características extraídos pela DRL VGG-Face, a implementação de uma rede LSTM foi capaz de modelar as informações temporais com apenas uma camada recorrente. E, apesar do ganho em performance não ser muito expressivo, a LSTM melhorou em 6,33% o resultado $max - min$ F_1-score entre as poses da cabeça, mostrando uma maior uniformidade entre a pior e melhor pose da base de dados.

Por último, a região dinâmica de aprendizado aplicada na VGG-Face combinada com modelagem temporal, superou o estado-da-arte na base de dados FERA 2017. Utilizando a combinação entre uma rede neural convolucional e uma rede neural recorrente para a predição *multi-label* de AUs, o método obteve um F_1-score médio de 0,593.

Em trabalhos futuros, pode-se ser aperfeiçoar a combinação entre a rede neural convolucional e a rede neural recorrente através de um treino unificado, diminuindo assim as chances de *overfitting*. Além disso, os passos seguintes incluem: o desenvolvimento de um método para combinar os AUs detectados, transformando-os em expressões faciais básicas e compostas; e a validação em ambientes não controlados.

Referências

[Albiero et al., 2018] Albiero, V., Bellon, O. R. P. e Silva, L. (2018). Multi-label action unit detection on multiple head poses with dynamic region learning. Em *ICIP*.

[Ali et al., 2017] Ali, A. M., Alkabbany, I., Farag, A., Bennett, I. e Farag, A. (2017). Facial action units detection under pose variations using deep regions learning. Em *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*.

[Baltrušaitis et al., 2015] Baltrušaitis, T., Mahmoud, M. e Robinson, P. (2015). Cross-dataset learning and person-specific normalisation for automatic action unit detection. Em *FG*.

[Batista et al., 2017] Batista, J. C., Albiero, V., Bellon, O. R. P. e Silva, L. (2017). Aumpnet: Simultaneous action units detection and intensity estimation on multipose facial images using a single convolutional neural network. Em *FG*.

[Benitez-Quiroz et al., 2016] Benitez-Quiroz, C. F., Srinivasan, R. e Martinez, A. M. (2016). Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. Em *CVPR*.

[Bulat e Tzimiropoulos, 2017] Bulat, A. e Tzimiropoulos, G. (2017). How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). Em *ICCV*.

[Chollet, 2017] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *arXiv preprint*.

[Chu et al., 2017] Chu, W.-S., De la Torre, F. e Cohn, J. F. (2017). Learning spatial and temporal cues for multi-label facial action unit detection. Em *FG*.

[Du et al., 2014] Du, S., Tao, Y. e Martinez, A. M. (2014). Compound facial expressions of emotion.

[Ekman et al., 2002] Ekman, P., Friesen, W. e Hager, J. (2002). *Facial Action Coding System (FACS): Manual*. A Human Face.

[He et al., 2017] He, J., Li, D., Yang, B., Cao, S., Sun, B. e Yu, L. (2017). Multi view facial action unit detection based on cnn and blstm-rnn. Em *FG*.

[He et al., 2016] He, K., Zhang, X., Ren, S. e Sun, J. (2016). Deep residual learning for image recognition. Em *CVPR*.

[Hochreiter e Schmidhuber, 1997] Hochreiter, S. e Schmidhuber, J. (1997). Long short-term memory. *Neural computation*.

[Jain e Learned-Miller, 2010] Jain, V. e Learned-Miller, E. (2010). Fddb: A benchmark for face detection in unconstrained settings. Relatório técnico, University of Massachusetts, Amherst.

[Li et al., 2017a] Li, W., Abtahi, F. e Zhu, Z. (2017a). Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. Em *CVPR*.

[Li et al., 2018] Li, W., Abtahi, F., Zhu, Z. e Yin, L. (2018). Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *TPAMI*.

[Li et al., 2017b] Li, X., Chen, S. e Jin, Q. (2017b). Facial Action Units Detection with Multi-Features and -AUs Fusion. Em *FG*.

[Martinez et al., 2017] Martinez, B., Valstar, M. F., Jiang, B. e Pantic, M. (2017). Automatic analysis of facial actions: A survey. *TAC*.

[Parkhi et al., 2015] Parkhi, O. M., Vedaldi, A. e Zisserman, A. (2015). Deep face recognition. Em *BMVC*.

[Ren et al., 2015] Ren, S., He, K., Girshick, R. e Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Em *NIPS*.

[Simon et al., 2016] Simon, M., Rodner, E. e Denzler, J. (2016). Imagenet pre-trained models with batch normalization. *arXiv:1612.01452*.

[Simonyan e Zisserman, 2015] Simonyan, K. e Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556*.

[Srivastava et al., 2014] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. e Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *JMLR*.

[Szegedy et al., 2016] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. e Wojna, Z. (2016). Rethinking the inception architecture for computer vision. Em *CVPR*.

[Tang et al., 2017] Tang, C., Zheng, W., Yan, J., Li, Q., Li, Y., Zhang, T. e Cui, Z. (2017). View-independent facial action unit detection. Em *FG*.

[Valstar et al., 2017] Valstar, M., Lozano, E. S., Cohn, J. F., Jeni, L. A., Girard, J. M., Yin, L., Zhang, Z. e Pantic, M. (2017). Fera 2017 - addressing head pose in the third facial expression recognition and analysis challenge. Em *FG*.

[Valstar et al., 2015] Valstar, M. F., Almaev, T., Girard, J. M., McKeown, G., Mehu, M., Yin, L., Pantic, M. e Cohn, J. F. (2015). Fera 2015 - second facial expression recognition and analysis challenge. Em *FG*.

[Yang et al., 2016] Yang, S., Luo, P., Loy, C. C. e Tang, X. (2016). Wider face: A face detection benchmark. Em *CVPR*.

[Zavan et al., 2017] Zavan, F. H. d. B., Gasparin, N., Batista, J. C., Silva, L. P. e., Albiero, V., Lucio, D. R., Bellon, O. R. e Silva, L. (2017). Face analysis in the wild. Em *SIBGRAPI)*.

[Zhang et al., 2014] Zhang, X., Yin, L., Cohn, J. F., Canavan, S., Reale, M., Horowitz, A., Liu, P. e Girard, J. M. (2014). Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *IMAVIS*.

[Zhao et al., 2016] Zhao, K., Chu, W.-S. e Zhang, H. (2016). Deep region and multi-label learning for facial action unit detection. Em *CVPR*.

[Zoltán et al., 2016] Zoltán, T., László, A. J., András, L. e Cohn, J. F. (2016). Deep learning for facial action unit detection under large head poses. Em *ECCV*.